Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Lexical variables that affect integrated speaking test scores among EFL learners

Hyojung Lim

lim@kw.ac.kr

Kwangwoon University

Research Article

Keywords: Integrated speaking tasks, Source text modality, Lexical sophistication, Lexical diversity, Lexical fluency, Korean English Learners' Spoken Corpus (KELSC)

Posted Date: July 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10094910/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

This study investigates the role of lexical richness in integrated speaking tasks among Korean EFL learners, focusing on how source text modality (aural vs. written) mediates the relationship between linguistic variables and oral proficiency. Drawing on the Korean English Learners' Spoken Corpus (KELSC), the study analyzed 474 transcribed oral responses from university students completing listen-to-speak and read-to-speak summary tasks evaluated on the CEFR scale. Automated text analysis tools were utilized to extract multi-dimensional indices of lexical sophistication, diversity, and fluency, alongside syntactic and cohesive controls. Ordinal logistic regression and partial proportional odds modeling (PPOM) revealed distinct patterns across the two task types. In the listen-to-speak task, lexical fluency (word count), diversity (MATTR), and sophistication (low spoken frequency and strong bigram association) collectively and linearly predicted higher CEFR levels. Conversely, in the read-to-speak task, lexical sophistication (Age of Acquisition) exerted a uniform positive effect across all levels, whereas lexical fluency and syntactic complexity only differentiated beginner-level users. Notably, lexical diversity exhibited an inverted-U pattern in the read-to-speak task, peaking at intermediate levels before declining among advanced speakers, who prioritized precise, succinct vocabulary over varied repetition. These findings suggest that written source texts can serve as a linguistic scaffold that alters the developmental trajectory of productive vocabulary. The study offers practical implications for designing task-specific scoring rubrics and tailoring lexical pedagogy to learners' specific proficiency bands in integrated assessment contexts.

Introduction

Speaking in an L2 requires multiple cognitive and linguistic competencies (Green, 2013). Speaking begins with the formulation and organization of ideas; if there is nothing to express in the first place, a learner cannot initiate the speaking process. Once ideas are formed, the learner moves to the encoding phase, where those ideas are converted into language internally. Learners with extensive knowledge of L2 vocabulary and grammar are presumably at an advantage because they have more resources to draw on, though the automatized use of that knowledge is a separate matter. After the ideas are successfully encoded, the learner must then draw on their knowledge of phonetics and phonology to make the intended language audible. We cannot rule out the role of metacognition, which plans, monitors, and remedies the entire process of speaking (Levelt, 1989). The current study aims to focus on the encoding phase, the primary stage where second language learners often encounter significant bottlenecks.

Theoretically and empirically, vocabulary knowledge has been one of the most reliable indicators of second language (L2) competence. Miralpeix and Muñoz (2018) revealed that receptive vocabulary size significantly predicted one's listening, reading, speaking and writing. Milton et al. (2010) claimed that vocabulary knowledge alone could explain 40 to 60 percent of the variance in four language skills tested in IELTS. The relationship between vocabulary knowledge and language skills appears self-evident at first glance, but closer examination reveals considerable complexity. Vocabulary knowledge is a

multidimensional construct (e.g., Nation, 2013), and its contribution to language ability can vary with the skill in question, whether written or spoken, and productive or receptive.

Speaking unfolds as an immediate and transient process, affected by temporal constraints and cognitive load, which results in restrictions on linguistic range and sophistication that speakers can use. In writing, proficient L2 writers typically employ low-frequency words, indicating accessibility to large mental lexicon (Kim et al., 2018; Monteiro et al., 2020); in speaking, however, the use of low-frequency words does not necessarily feature advanced L2 speaking proficiency (Eguchi & Kyle, 2020). Instead, proficient L2 speakers tend to use high-frequency words with fluency and flexibility (Eguchi & Kyle, 2020; Kao & Wang, 2014). Task types also mediate the relationship between vocabulary and productive skills. Sophisticated vocabulary carries greater weight in argumentative essay writing compared to narrative and expository essay writing, in both of which fluency emerges as the strongest predictor of the test scores (Yang et al., 2025). Similarly, lexical sophistication was the most important predictor of independent writing tasks, which was not the case in the integrated writing tasks (Kyle & Crossley, 2015). In speaking, the contribution of vocabulary knowledge to L2 speaking test scores appears larger in the integrated task than in the independent task (Appel et al., 2019).

Taken together, the present study aims to investigate the role of vocabulary in integrated speaking tasks among EFL learners. Drawing on the Korean English Learners' Spoken Corpus (Jung, 2021), which comprises one-minute oral responses from Korean university students completing both listen-to-speak and read-to-speak test formats, the study zooms in on learners’ speaking performance on each task type separately to elucidate the influence of source modality.

Lexical richness

Lexical richness has been identified as a key determinant of productive language skills. The umbrella term lexical richness usually refers to lexical sophistication, diversity, and density (Appel et al., 2019; Ha, 2019; Lu, 2012), encompassing the multiple dimensions of lexical competence. In Kolahi Ahari et al. (2025), both lexical sophistication and diversity accounted for 34% of variance in the descriptive, narrative monologue speaking performance among Iranian EFL learners, when combined with syntactic sophistication, complexity, and cohesion. In Sung et al. (2025), lexical richness, operationalized through lexical sophistication, diversity, and density, significantly predicted TOEFL junior writing scores.

Lexical sophistication (LS) means learners’ use of advanced or challenging vocabulary, traditionally conceived as low-frequency words. LS has been quantified through mean word frequency scores. In writing, it is evident that advanced learners can employ infrequent and rare words (Kim et al., 2018; Monteiro et al., 2020), as they are assumed to have larger mental lexicon than less proficient learners. Note that writing genres can serve as a critical factor; In Kyle and Crossley (2016), lexical sophistication outweighed diversity in predicting independent, opinion-based writing scores, whereas the reverse held true for source-based, integrated writing tasks. Speaking tasks present a more complex picture. Several studies found that more proficient speakers actually use higher-frequency words (Eguchi & Kyle, 2020;

Crossley & McNamara, 2013), presumably indicating learners' awareness of spoken registers. Other research, however, maintains the traditional association between lower-frequency words and higher speaking scores, especially on persuasive speech (Kim, 2024).

More recently, applied linguists have reconceptualized LS as multidimensional, highlighting diverse psycholinguistic lexical properties. Namely, the proportion of less concrete, less imageable, less familiar, and less semantically rich words signals lexical as well as cognitive sophistication (Kim et al., 2018; Kyle & Crossley, 2015). Lexical proficiency is also indicated when learners use words acquired later in native speaker development (Crossley & McNamara, 2012). In essence, proficient learners are expected to command cognitively demanding vocabulary and deploy sophisticated words while speaking and writing. The psycholinguistic features are also sensitive to both modality and genre. In Heng et al. (2023), Chinese EFL learners used less familiar and less meaningful words in expository writing than in argumentative writing. In Kyle et al. (2016), learners used words with higher meaningfulness values in the integrated speaking tasks than in the independent tasks.

Lexical diversity (LD) measures the range of distinct words used in a text, indicating a learner's productive vocabulary breadth. It also implies a learner's awareness of register, part of *Use* knowledge in Nation's (2013) term, because proficient learners should be able to flexibly choose a word considering a context. In general, written texts exhibit higher LD than spoken discourse (Vasylets et al., 2017), since the asynchronous nature permits more deliberate and comprehensive lexical retrieval. Nevertheless, lexical diversity has consistently emerged as a robust predictor of L2 speaking proficiency (Kolahi Ahari et al., 2025; Lu, 2012). Both native and trained raters appear to recognize varied word use as indicative of L2 speaking competence, associating it with enhanced comprehensibility and nativelikeness. The predictive strength of LD, however, varies across proficiency levels. Kolahi Ahari et al. (2025) found that LD was the strongest predictor for independent users (B1-B2 levels), while among proficient users (C1-C2 levels), it ranked third in importance behind cohesion measures.

Numerous metrics have been developed to quantify lexical diversity (LD) accurately. These indices fall into two categories: global measures that analyze the entire text simultaneously, and local measures that examine sequential segments (Bestgen, 2024). As a global index, the type-token-ratio (TTR), the proportion of unique words to total words, represents the simplest and most intuitive measure. However, its susceptibility to text length effects has prompted various mathematical refinements, including Guiraud's R, Maas's a, Root TTR, and Hypergeometric Distribution (HD-D). Local indices, on the other hand, address sample size limitations through segmented analysis rather than whole-text calculation. Mean segmental TTR (MSTTR) averages TTR across fixed, non-overlapping blocks, though this approach discards end-of-text data. Moving-Average TTR (MATTR) employs a sliding window that incorporates every word, while Measure of Textual Lexical Diversity (MTLD) measures the average text length required to reach a specified TTR threshold. Through extensive empirical review, Bestgen recommends HD-D for global analysis, and MATTR or MTLD for local analysis. For the analysis of short texts, MATTR and MTLD have demonstrated validity and reliability in both L2 writing (Koizumi & In'nami, 2012) and L2 speaking (Kyle et al., 2024)

Two additional lexical variables frequently examined are lexical density and fluency. Lexical density represents the proportion of content words to total words (Ure, 1971), conceptually indicating how much informational content is concentrated within a text. Its relationship with L2 writing quality remains contested: some researchers found no significant association (Engber, 1995; Linnarud, 1986), while Ha (2019) reported a negative correlation, proposing that excessive content word density might reflect limited grammatical competence. Other studies, however, have documented positive relationships (Kim, 2014; Gregori-Signes & Clavel-Arroitia, 2015). In speaking, lexical density is usually lower than in writing because real-time speech production relies more on function words to maintain fluency and manage interaction. As with writing, lexical density has not gained universal acceptance as a robust predictor of oral proficiency (Lu, 2012). Task interactivity significantly influences lexical density (O'Loughlin, 1995): highly interactive tasks such as live role-plays tend to yield lower lexical density compared to less interactive formats like recorded monologues.

Lexical fluency indicates the speed and automaticity with which learners can access and retrieve words from their mental lexicon (Dóczi & Kormos, 2015). It is identified as vocabulary depth knowledge (Schmitt & Schmitt, 2020) or part of lexical richness (Ha, 2019). It is often measured as the total number of words produced within a given time. Advanced L2 learners generally demonstrate both rapid and accurate vocabulary use in written and spoken contexts (Cumming et al., 2005; Uchihara & Clenton, 2023). Notably, in spoken production, fillers, pauses, and repetitions are indicative of reduced fluency. Lexical fluency also contributes indirectly to language proficiency by enabling efficient word retrieval, which releases cognitive resources for other linguistic or non-linguistic dimensions of communication, such as syntactic processing or content development (Zhang & Lu, 2014). Note that vocabulary size, encompassing both receptive and productive knowledge, shows positive correlations with oral fluency (de Jong & Mora, 2019; Uchihara & Saito, 2019). In this regard, isolating the specific contribution of lexical fluency to language proficiency proves challenging due to its complex interconnections with other variables. As with other lexical measures, the effect of fluency varies across genres or task types, showing considerably stronger predictive validity for narrative and expository tasks than for argumentative tasks (Yang et al., 2025).

Collocation use, as operationalized in n-grams, is strongly associated with lexical fluency. For instance, trigram (three-word sequence) frequency emerged as the most powerful predictor of independent speaking scores (Kyle & Crossley, 2015). Kim et al. (2018) also revealed that ESL learners produced increasingly common bigrams and trigrams, along with more frequent bigrams, over one year of instruction. While collocation knowledge influences both accuracy and fluency in language production (Laufer & Waldman, 2011), English language learners are likely to have weaker collocation knowledge compared to native speakers (Webb & Nation, 2017). Consequently, proficient use of L2 collocation can effectively indicate advanced language proficiency.

Other linguistic variables

Syntactic knowledge constitutes another fundamental pillar of L2 proficiency. Some scholars argue for a lexicogrammatical perspective, claiming that lexical items inherently encode grammatical information, rendering vocabulary versus grammar separation artificial (Römer, 2009). However, componential models (Khalifa & Weir, 2009) and neurocognitive evidence (Kimppa et al., 2019) support distinguishing syntactic from lexical knowledge. Empirically, these two are typically treated as separate yet strongly correlated factors that jointly shape language proficiency (Cai, 2020; Vafaee & Suzuki, 2020).

The methodological challenge in L2 writing and speaking research involves developing reliable measures of syntactic knowledge. While traditional metrics like mean length of T-unit (MLTU) correlate with writing quality, fine-grained phrasal complexity indices show stronger predictive power for L2 writing proficiency than clausal measures (Kyle & Crossley, 2018). Noun phrase modification, such as adjectives and prepositional phrases, accounts for substantial variance in essay scores, corroborating that academic writing development involves progression from clausal subordination toward phrasal elaboration.

Syntactic complexity manifests differently in L2 speaking, characterized primarily by finite subordination such as dependent clauses per clause rather than the phrasal embedding characteristic of written discourse (Kim, 2024). Moreover, syntactic knowledge's predictive power varies across proficiency levels. Kolahi Ahari et al. (2025) found that syntactic sophistication, operationalized as advanced verb-argument construction usage, strongly predicted speaking proficiency among advanced learners (C1-C2) but showed no predictive value for independent-level users (B1-B2). Among lower-proficiency learners, lexical diversity and accuracy emerged as stronger predictors than syntactic features.

Learners' use of cohesive devices and coherence-building strategies also enhances both L2 writing and speaking quality. In theory, effective discourse should maintain coherence through organizing supporting details around central ideas. However, L2 writing research reveals a complex relationship where increased cohesion does not necessarily contribute to quality. Advanced L2 writers often produce texts with fewer explicit cohesive devices, either anticipating reader inference or employing more sophisticated implicit connection strategies (Crossley & McNamara, 2012). Similarly, in L2 speaking, coherence, meaning the logical linking of ideas, correlates strongly with comprehensibility, though explicit cohesive devices contribute variably across proficiency levels (Tsunemoto & Trofimovich, 2024). Mechanical overuse of such devices can harm ratings, especially at intermediate levels, while advanced performance prioritizes implicit coherence and sophisticated syntax over explicit cohesive markers. Regarding referential cohesion, Crosthwaite (2011) documented differences between native and non-native speakers; native English speakers often employ personal pronouns for co-reference and definite articles for distant reference to maintain textual cohesion. Korean learners of English, however, struggled to deploy these devices in native-like patterns, instead overusing bare nominals or repetition, which undermined narrative coherence.

Integrated tasks

Task type mediates the relationship between linguistic variables and L2 proficiency. In Crowther et al. (2015), a wider range of linguistic variables (lexis, grammar, and discourse) collectively predicted the comprehensibility of the TOEFL iBT integrated speaking task, compared to the IELTS long-turn independent speaking task, in which only pronunciation and fluency emerged as significant predictors. Learners' responses to integrated writing tasks also show more academic lexicogrammatical features than their responses to independent tasks, such as passive voice, nominalization, longer T-units and more clauses (Kyle, 2020). That is, integrated tasks seem to successfully elicit more complex grammar structures as well as more sophisticated vocabulary use. Learners' use of cohesive devices has a differing impact on their speaking performance across task types. Advanced L2 writers write and speak economically in independent tasks, using fewer explicit connectives (Crossley & McNamara, 2012), whereas the increased use of local cohesive devices gives rise to higher scores in integrated tasks (Guo et al., 2013). That being said, when coherence is measured by semantic overlap, independent tasks in effect demonstrate significantly higher coherence values than integrated tasks (Ariamanesh et al., 2022).

The link between lexical richness and the integrated writing scores appears somewhat complex, as source texts constrain as well as support production. Source texts can reduce cognitive planning demands, enabling learners to produce writing with enhanced lexical accuracy, richness, and sophistication (Abrams, 2019). Conversely, other studies indicate that integrated tasks could yield reduced lexical productivity compared to independent tasks, presumably because the integrated format constrains the expected vocabulary range (Kyle et al., 2016; Sung et al., 2025). In integrated tasks, learners who successfully incorporated frequently repeated keywords from source texts tend to achieve higher proficiency scores (Crossley et al., 2014), which in fact results in reduced lexical diversity. Ironically, as a result of recycling words from written source texts, oral responses to integrated academic tasks likely contain more sophisticated vocabulary, or less frequent words in spoken corpora. (Kyle et al., 2016; Kyle, 2020). Notably, learners are likely to make more linguistic errors in the integrated tasks than in the independent task, as they attempt to use more complex vocabulary and grammar on relatively unfamiliar topics (Ariamanesh et al., 2022).

The modality of source texts also impacts how raters interpret learners' use of source materials. Kyle (2020) reported that integrating content words and n-grams from listening passages correlated positively with writing test scores, whereas excessive reliance on reading passages often correlated negatively with the test scores. The capacity to recall and integrate auditory information seems to effectively identify high-proficiency learners, while simply copying sophisticated vocabulary from written texts could rather evidence a limited vocabulary knowledge.

Although source-text modality has been shown to reshape the relationship between lexical features and proficiency in integrated writing, its role in integrated speaking remains largely unexplored, as research on lexical richness has concentrated on writing and on independent speaking tasks. It thus remains unclear whether the lexical features that signal speaking proficiency, and the way they develop across levels, are the same when learners summarize what they have heard as when they summarize what they

have read. The present study addresses this gap by modeling how lexical richness predicts L2 speaking proficiency separately in listen-to-speak and read-to-speak tasks, drawing on a corpus of Korean EFL learners' speech elicited through locally developed integrated tasks rather than a standardized proficiency test (KELSC). The current study is guided by the following research questions:

1. To what extent does lexical richness affect L2 speaking scores in the listen-to-speak task?
2. To what extent does lexical richness affect L2 speaking scores in the read-to-speak task?

Methodology

The Korean English Learners' Spoken Corpus (KELSC)

The KELSC is the publicly available corpus of spoken English produced by Korean adult learners of English (Jung, 2021). Constructed from real-time online video recordings of 237 EFL university students, the corpus consists of 474 transcribed texts totaling 36,588 tokens and 1,799 types. The participants performed integrated speaking tasks without the aid of dictionaries or the Internet, and their outputs were evaluated based on the Common European Framework of Reference for Languages (CEFR) scale ranging from A1 to C2.

The KELSC includes two specific types of integrated speaking tasks: listening-to-summarize and reading-to-summarize, which were administered in two sets (Type A and Type B). The Type A tasks required participants to listen to a school announcement regarding a change in the schedule for a sports day and summarize it, and to read a passage about successful fundraising examples using social media and summarize the content. The Type B tasks involved listening to an announcement regarding apartment remodeling and reading a text about the role of senior welfare centers in an aging society and summarizing the information. For each task, participants were generally allotted one minute to listen or read, one minute to prepare, and one minute to record their spoken response (For a detailed overview of the corpus, see Jung, 2021). Table 1 presents the number of transcribed spoken texts across CEFR levels along with mean speech length, measured as the number of words produced.

Table 1
Description of KELSC

		A1	A2	B1	B2	C1	C2
Listen-to-speak tasks	Number of texts	12	30	81	59	40	13
	Word Count Mean	23.08	43.90	60.53	73.61	75.88	87.38
Read-to-speak tasks	Number of texts	12	30	81	59	40	13
	Word Count Mean	30.33	52.90	70.72	82.44	85.53	107.08

Text preprocessing

The KELSC corpus was preprocessed before the subsequent textual analysis. For lexical analysis, the following elements were removed from the transcribed texts: (1) filler words (e.g., ah, ahm, um, hmm), regardless of case; (2) incomplete words resulting from stammering (e.g., succ... for success); (3) ellipses (...); (4) punctuation marks (periods and commas); (5) extraneous spaces; and (6) word repetitions (e.g., the the) arising as artifacts of the removal processes described in steps 1–5. For syntactic and cohesion analyses, however, punctuation was retained while all other preprocessing procedures remained identical.

Lexical sophistication

To quantify lexical sophistication, the pre-processed texts were submitted to the TAALES 2.0 toolkit (Kyle et al., 2018). Lexical sophistication was operationalized in two ways. First, averaged word frequency was calculated representing the mean frequency of words within each text. Both COCA spoken and academic frequency values were considered. Lexical sophistication indices for content words were selected over those computed for all words, as content words more directly reflect productive vocabulary knowledge. Moreover, log-transformed frequency indices were used instead of raw frequency values (Kyle & Crossley, 2015). Second, psycholinguistic word properties served as another set of lexical sophistication measures, including lexical familiarity, concreteness, imageability, meaningfulness, and age of acquisition (AoA). It is assumed that more proficient learners would be able to employ more cognitively demanding vocabulary; that is, words with lower frequency, reduced familiarity, less concrete, and later acquisition ages.

Lexical diversity

Both MATTR and MTLD indices were utilized to measure lexical diversity (LD), calculated using the Python *lexical diversity* library. Of the 474 analyzed texts, four (two from each task type) were excluded because they contained fewer than the 11 tokens required for MATTR. As shown in Table 2, MATTR scores for listen-to-speak tasks increased modestly with L2 proficiency, alongside a decrease in within-group variation. However, MATTR scores for read-to-speak tasks seemed comparable across all six levels. Given that the MATTR values ranged between 0.88 and 0.95, participants demonstrated a relatively high level of lexical diversity compared to prior studies. Note that in the TOEFL Junior Writing test, students receiving the highest proficiency scores (Level 4) exhibited mean MATTR values of 0.953 for Opinion tasks and 0.856 for Listen-Write tasks (Sung et al., 2025). Meanwhile, a considerably larger number of texts (n = 117) failed to produce MTLD values, as this measure requires a minimum of 50 tokens. Therefore, only MATTR indices were used in the subsequent regression analyses.

Table 2
MATTR values across CEFR levels

A1	Listen-to-speak tasks	Read-to-speak tasks
	0.88(0.07)	**0.93(0.08)**
A2	0.88(0.06)	0.92(0.05)
B1	0.91(0.04)	0.93(0.04)
B2	0.92(0.04)	0.94(0.04)
C1	0.92(0.03)	0.95(0.02)
C2	0.94(0.02)	0.93(0.04)

Other lexical variables

The indices for syntactic complexity and cohesion were collected by means of the automated analysis tools, Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC; Kyle, 2016) and the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley et al., 2016).

TAASSC computes hundreds of indices capturing syntactic development through both traditional metrics and detailed measures of clausal and phrasal complexity. Complexity is quantified through structural elaboration, such as mean length of units, and the frequency of dependents associated with particular syntactic structures. Sophistication is evaluated by analyzing the frequency and association strength of verb-argument constructions relative to five sub-corpora from the COCA. This study opted for traditional metrics, using 14 holistic measures (Lu, 2010); first, automated natural language processing tools generally show higher reliability when computing holistic measures than fine-grained ones (Kim, 2024). Second, the current study aims to control the effect of syntactic complexity on speaking test scores rather than pinpointing isolated structural occurrences. Third, the small set of traditional metrics can help avoid severe multicollinearity and model overfitting in subsequent regression analyses.

TAACO generates several hundred indices for local, global, and overall text cohesion measures. It reports various types of connectives based on Halliday and Hasan (1976), including logical and additive connectives (e.g., *and, moreover*), adversative and contrastive connectives (e.g., *but, however*), causal and temporal connectives (e.g., *because, first, until*), and disjunctive connectives (e.g., *or*). It also calculates the frequency with which words and ideas recur across different text segments, such as lexical overlap and synonym overlap. The present study used the most updated version TAACO 2.0., allowing for Latent Semantic Analysis (LSA), Word2vec, and Latent Dirichlet Allocation (LDA).

Data analysis

To manage the extensive output of the automated text analysis tools, the number of independent variables was systematically reduced through multiple processes, based on previous findings (Crowther et al., 2015; Sung et al., 2025; Tsunemoto & Trofimovich, 2024) as well as a theoretical definition of L2 speaking ability (Green, 2013). To minimize multicollinearity, indices that capture similar constructs were screened based on correlation analyses; for example, *frequency* was retained over *range* ($r > .85$), and *imageability* over *concreteness* ($r > .90$). Taken together, the number of regressors came down to 22, with the MATTR values included.

To identify predictors of CEFR proficiency levels, ordinal LASSO regression (L1-penalized cumulative logit model) was applied to 22 candidate linguistic variables using the *ordinalNet* package in R (Wurm et al., 2021) for each task. All predictors were standardized prior to analysis to ensure comparability of coefficients. The regularization parameter λ was selected via 10-fold cross-validation using the 1 standard error rule (λ.1se; Tibshirani, 1996), which identifies the most parsimonious model whose cross-validated misclassification rate falls within one standard error of the minimum. Following LASSO selection, variance inflation factors (VIF) were examined to detect multicollinearity, with a threshold set at VIF < 5 (Liu et al., 2023). Consequently, the listen-to-speak task ended up with 14 variables while the read-to-speak task with nine variables.

To assess the proportional odds assumption, both the Brant test (Brant, 1990) and the likelihood ratio-based nominal test (Christensen, 2019) were performed from the ordinal package. The Brant omnibus test for the listen-to-speak task was not significant ($\chi^2 = 52.24$, $df = 56$, $p = .618$), thereby supporting a standard cumulative logit model for the listen-to-speak data. Meanwhile, the Brant test for the read-to-speak task turned out to be significant ($\chi^2 = 89.35$, $df = 36$, $p < .001$). Three variables were identified with non-proportional effects: MATTR ($p = .002$), *Word Count* ($p = .014$), and *C/S* ($p = .026$). The likelihood ratio-based nominal test corroborated these findings, with the same three variables exhibiting significant departures from proportionality. A partial proportional odds model (Peterson & Harrell, 1990) was therefore adopted for the read-to-speak task, allowing *Word Count*, *C/S*, and *MATTR* to have threshold-specific coefficients while maintaining proportional effects for the remaining six predictors.

Results

1. *Lexical effects on the listen-to-speak task*

Based on the LASSO, VIF analysis, and Brant test results, 14 predictors were entered into the ordinal logistic regression. The model significantly improved over the null (*LR* $\chi^2 = 123.30$, $df = 14$, $p < .001$), with a *Nagelkerke* R^2 of .426, indicating that the predictors accounted for approximately 43% of the variance in CEFR proficiency levels. The model correctly classified 44.3% of cases exactly and 87.7% within one adjacent CEFR level.

Four of the 14 LASSO-selected predictors reached statistical significance in the ordinal logistic regression (Table 3). *Word Count* was the strongest predictor ($B = 1.09$, *OR* = 2.98, *95% CI* [2.14, 4.20], $p < .001$): a one standard deviation increase in word count was associated with nearly three times higher

odds of being classified at a higher CEFR level. *MATTR* (B = 0.41, *OR* = 1.51, *95% CI* [1.16, 1.97], *p* = .002) showed that greater lexical diversity was associated with 51% higher odds of a higher proficiency rating. *COCA Spoken Bigram MI*^{2} (B = 0.39, *OR* = 1.47, *95% CI* [1.09, 2.00], *p* = .013) suggested that stronger collocational association strength in spoken registers was positively associated with proficiency. Conversely, *COCA Spoken Frequency (Log)* showed a negative association (B = -0.36, *OR* = 0.70, *95% CI* [0.51, 0.96], *p* = .028), indicating that the use of less frequent spoken vocabulary was associated with higher CEFR levels. Finally, the contribution of *Basic Connectives* approached significance at the .05 level (B = -0.31, *OR* = 0.733, *95% CI* [0.53, 1.01], *p* = .055). The negative coefficient suggests a marginal trend where a lower reliance on basic connectives goes hand in hand with higher CEFR proficiency levels.

Table 3
Results of the ordinal logistic regression for the listen-to-speak task

Variable	***B***	***SE***	***z***	**OR**	**95% CI**	***p***
MATTR**	0.41	0.13	3.04	1.505	[1.16, 1.97]	.002
Word Count***	1.09	0.17	6.35	2.981	[2.14, 4.20]	< .001
MRC Imageability	−0.26	0.17	−1.56	0.767	[0.55, 1.07]	.119
COCA Spoken Freq. (Log)*	−0.36	0.16	−2.19	0.701	[0.51, 0.96]	.028
USF	−0.02	0.15	−0.16	0.976	[0.72, 1.31]	.874
COCA Spoken Bigram MI^2*	0.39	0.16	2.48	1.471	[1.09, 2.00]	.013
COCA Academic Bigram DP	−0.25	0.15	−1.70	0.779	[0.58, 1.04]	.089
Types per Family	0.13	0.14	0.91	1.140	[0.86, 1.52]	.363
Content Word %	−0.12	0.14	−0.87	0.884	[0.67, 1.17]	.386
MLT	0.04	0.23	0.19	1.045	[0.67, 1.65]	.846
MLC	0.08	0.14	0.62	1.088	[0.83, 1.43]	.538
C/S	0.24	0.23	1.03	1.267	[0.81, 2.00]	.304
All Connectives	0.23	0.17	1.36	1.254	[0.91, 1.74]	.173
Basic Connectives	−0.31	0.16	−1.92	0.733	[0.53, 1.01]	.055

Note. Standardized estimates (per 1 SD increase). OR = odds ratio; CI = confidence interval; USF = free association stimuli elicited content words; MLT = mean length of T-units; MLC = mean length of clause; C/S = clauses per sentence. * p < .05, ** p < .01, *** p < .001.

2. *Lexical effects on the read-to-speak task*

Given that the read-to-speak data violates the proportional odds assumption, nine variables were submitted to the partial proportional odds model (PPOM), which significantly improved model fit

compared to the standard ordinal logistic regression (*LR* χ^2 = 42.08, *df* = 12, *p* < .001). Among the six variables constrained to have uniform effects across all CEFR thresholds (Table 4), only *Kuperman Age of Acquisition* reached statistical significance (*B* = 0.34, *OR* = 1.40, *p* = .038). This indicates that the use of later-acquired, more sophisticated vocabulary was consistently associated with higher CEFR proficiency across all levels. Additionally, *COCA Academic Bigram DP* seemed to approach significance (*p* = .072), suggesting a marginal role for collocational knowledge in academic registers.

Table 4
PPOM proportional coefficients (uniform Effects across all thresholds)

Variable	***B***	***SE***	***z***	**OR**	**95% CI**	***p***
COCA Academic Freq. (Log)	−0.27	0.22	−1.24	0.77	[0.50, 1.17]	.217
MRC Imageability	−0.21	0.14	−1.53	0.81	[0.62, 1.06]	.125
COCA Spoken Freq. (Log)	0.02	0.24	0.08	1.02	[0.64, 1.62]	.935
Kuperman AoA*	0.34	0.16	2.07	1.40	[1.02, 1.92]	.038
COCA Academic Bigram DP	0.24	0.13	1.80	1.27	[0.98, 1.64]	.072
Content Word %	−0.16	0.15	−1.10	0.85	[0.64, 1.13]	.272

Note. Standardized estimates (per 1 SD increase). OR = odds ratio; CI = confidence interval. These variables were constrained to have equal effects across all CEFR thresholds based on non-significant Brant individual tests. * p < .05.

With regard to the three variables (*Word Count*, *MATTR*, *C/S*) identified as non-proportional effects, a series of binary logistic regressions was conducted at each CEFR threshold (Table 5). Each variable shows qualitatively distinct developmental patterns (Table 5). *Word Count* showed a monotonically decreasing pattern across thresholds. The effect was strongest at the lowest threshold (CEFR ≥ 2 vs. ≤ 1; *OR* = 54.02), declined substantially at the lower-intermediate level (*OR* = 6.35), and stabilized at moderate levels for upper thresholds (*OR* = 2.09–2.73). This pattern indicates that lexical fluency could be a determining factor at the lowest proficiency levels; learners at the A1 level produce markedly less speech in the read-to-speak task. However, it becomes a less powerful predictor as proficiency increases, where qualitative aspects of production may carry greater weight.

C/S (Clauses per Sentence) displayed a steep decline to near-null effect. At the lowest threshold, the *OR* was extremely large (*OR* = 178.08), though this estimate should be interpreted with caution due to quasi-complete separation (n = 12 for A1). The effect diminished rapidly across subsequent thresholds (*OR* = 4.10, 1.61, 1.96) and was virtually absent at the highest threshold (*OR* = 1.03). Syntactic complexity, as measured by clausal embedding, seems to effectively differentiate lower proficiency levels but rarely contribute to distinguishing advanced from the most advanced speakers.

Finally, MATTR demonstrated a qualitatively distinct *inverted-U pattern* with direction reversal. The effect was weak at the lowest threshold (*OR* = 1.23), increased to its peak at the intermediate levels (CEFR ≥ 4

vs. ≤ 3; *OR* = 1.53; CEFR ≥ 5 vs. ≤ 4; OR = 1.50), and reversed direction at the highest threshold (*OR* = 0.86). While greater lexical diversity facilitates progression through intermediate CEFR levels, the most advanced speakers at C2 seem to employ more focused, precise vocabulary rather than maximizing lexical diversity.

Table 5
Threshold-specific odds ratios from binary logistic regressions

Threshold	Word Count		C/S		MATTR	
	β	OR	β	OR	β	OR
CEFR ≥ 2 vs. ≤ 1	+ 3.989	54.017	+ 5.182	178.080	+ 0.208	1.231
CEFR ≥ 3 vs. ≤ 2	+ 1.849	6.351	+ 1.412	4.103	+ 0.326	1.385
CEFR ≥ 4 vs. ≤ 3	+ 0.965	2.624	+ 0.476	1.610	+ 0.427	1.532
CEFR ≥ 5 vs. ≤ 4	+ 0.739	2.093	+ 0.671	1.956	+ 0.402	1.495
CEFR ≥ 6 vs. ≤ 5	+ 1.004	2.728	+ 0.032	1.033	−0.154	0.857

Note. C/S = clauses per sentence; MATTR = moving-average type–token ratio; OR = odds ratio

Discussion

In both integrated tasks, lexical variables significantly predicted L2 speaking proficiency among Korean EFL learners. In particular, learners' lexical fluency, diversity, and sophistication all together influenced their listen-to-speak test scores, whereas lexical sophistication alone affected their read-to-speak test scores. The way in which the subcomponents of lexical competence affect learners' speaking performance in the read-to-speak test appears to be more nuanced, presumably reflecting a distinct developmental trajectory for each component of lexical knowledge.

In the listen-to-speak task, *Word Count, MATTR, COCA Spoken Frequency, and COCA Spoken Bigram MI*^{2} emerged as the significant predictors of CEFR band scores. Notably, *Word Coun*t serves as the most powerful predictor, indicating that lexical fluency played a decisive role in learners' speaking test performance. Also, *COCA Spoken Bigram MI*^{2} was positively associated with proficiency, suggesting that advanced learners produced more strongly associated word combinations typical of spoken register. The use of collocational and formulaic expressions helps facilitate fluent performance by allowing speakers to produce longer, uninterrupted speech without needing to encode each lexical and grammatical element (Eguchi & Kyle, 2020). In this regard, advanced learners' greater use of bigram units may signal enhanced fluency.

Moreover, *MATTR*, an indicator of lexical diversity, was likewise a significant positive predictor. Advanced learners were able to draw on a wider range of vocabulary when responding to the listen-to-speak task. The use of more varied lexical items reflects broader vocabulary knowledge and the ability to access

diverse forms under real-time production pressure (Crossley & McNamara, 2012). The listen-to-speak tasks seem to effectively identify learners with a higher degree of lexical fluency and diversity; compared to written source texts, aural input is inherently transient. It therefore imposes relatively higher cognitive load on learners, who must allocate considerable working memory resources to understanding the source text. Given that learners cannot control the pacing or replay the auditory input, lower-proficiency learners seem to experience cognitive overload during the listening phase, subsequently leading to frequent hesitation while recalling what they heard. Likewise, they would not have extra capacity to elicit various words from their mental lexicon, coupled with a baseline deficiency in vocabulary knowledge.

Finally, COCA Spoken Frequency was negatively associated with the CEFR scores, indicating that advanced learners used lower-frequency words while speaking. This contradicts previous findings, such as Eguchi and Kyle (2020), where proficient learners actually produced highly frequent content words in picture description and self-introduction tasks. In independent speaking tasks, the same trend was observed (Bottini, 2022; Bottini & Le Foll, 2025; Kyle & Crossley, 2015). However, these discrepancies may stem from methodological differences and task effects; Bottini's work did not use log-transformed word frequency values, whereas Kyle and Crossley (2015) utilized a mixed spoken and written corpus as a reference. More importantly, whereas participants in Eguchi and Kyle (2020) described pictures and comics, learners in the present study were exposed to an aural source text. Despite the ephemeral nature of aural input, learners who were able to uptake and reuse lower-frequency words seemed to obtain higher proficiency scores.

In the read-to-speak task, Kuperman AoA exerted a consistent influence across the entire spectrum of learner proficiency; higher-level learners naturally integrated words with a later age of acquisition. Much like in the listen-to-speak task, more proficient learners were able to use more sophisticated words when summarizing what they read. However, the unique predictive power of AoA seems to underscore an interaction between modality and task topics. While the listen-to-speak task featured a more conversational topic (e.g., apartment remodeling), the read-to-speak task introduced a highly abstract, expository theme (e.g., the role of senior welfare centers in an aging society). Given that written texts are likely to contain more academic words, spoken frequency might have been limited in distinguishing learners at higher proficiency levels. In contrast, Kuperman AoA spans a wider range of vocabulary, thereby effectively differentiating even between intermediate and advanced learners within this literate register. Those who could cognitively digest such sophisticated textual prompts and successfully recycle them during oral production seemed to earn higher proficiency scores.

Unlike in the listen-to-speak task, lexical fluency and diversity did not emerge as overarching predictors of speaking scores. Further scrutiny, however, reveals that the linguistic features underlying L2 oral proficiency operate differently across CEFR levels. Lexical fluency, operationalized as total word count, mattered only among beginning-level learners, with its predictive power diminishing once learners reached the intermediate level. In the read-to-speak task, where the written passage served as a content scaffold for subsequent oral production, no significant difference in the quantity of oral output was observed between B- and C-level learners. Despite the presence of visual support, beginning learners still

exhibited difficulty retrieving those words in the subsequent oral production. A parallel pattern was observed for syntactic complexity, measured as the number of clauses per sentence: this variable successfully distinguished learners only at the beginning level, while intermediate and advanced learners demonstrated comparable syntactic complexity, presumably attributable to the leveling effect of visual scaffolding. Note that none of the syntactic complexity measures (MLT, MLC, C/S) played a role in the listen-to-speak task, presumably for a different reason. In the listen-to-speak task, the cognitive demands of aural comprehension seem to constrain syntactic planning, rendering lexical and fluency-related measures more discriminative than structural complexity.

Finally, lexical diversity yielded an interesting distributional pattern. Diversity increased progressively from beginning to intermediate levels, yet declined among advanced learners. Possible scenarios are as follows: First, proficient learners might have prioritized efficiency in their oral delivery, opting for more sophisticated words rather than iterating across semantically similar terms. Second, stronger reading comprehension among advanced learners could have engendered more robust lexical priming, leading them to adhere closely to the key words in the source text. Third, intermediate learners may have compensated for gaps in their lexical repertoire by describing a single concept with several simple words, whereas advanced learners, having access to the precise lexical item, use the word once and move on. Such sophisticated yet succinct expressions may have paradoxically reduced overall lexical diversity. Taken together, when a written source text is provided and functions effectively as a vocabulary scaffold for the following speaking task, the contribution of linguistic variables to speaking scores seems to differ across proficiency levels.

Conclusion

In summary, depending on the modality of a source text, the effect of linguistic variables on integrated speaking scores turned out to vary. In the listen-to-speak task, lexical sophistication, diversity, and fluency collectively contributed to the speaking scores, while the effect of syntactic complexity and cohesion devices was not observed. Note that the use of sophisticated lexical items was assumed as proficient learners produced lower-frequency words in COCA spoken corpus. In the read-to-speak task, however, it was only lexical sophistication that uniformly affected the test scores; more proficient learners were able to use more cognitively demanding words. When the source text was written, lexical fluency and syntactic complexity in learners' oral production effectively identified beginning learners, but not upper levels of learners. More interestingly, lexical diversity increased up to intermediate levels but rather declined among advanced learners.

However, these findings should be interpreted with caution. The study is constrained by an uneven distribution of participants across CEFR levels: most learners clustered around the B1 and B2 levels, so the predictive power may diminish at the extremes of the proficiency spectrum, particularly in the lower bands. Although this distribution could be ecologically valid, as human ability tends to follow a normal distribution in populations, such unevenness can still affect statistical results. Also, the model's scope is limited by the exclusion of several critical variables. Specifically, the analysis does not account for

acoustic or phonological features of words, which are known to significantly influence oral proficiency ratings. Finally, the linguistic alignment between source texts and learners' oral output, which was beyond the scope of the current study, warrants further investigation.

The findings carry practical implications for both testing and teaching. For testing, a single rubric may be insufficient for scoring integrated speaking tasks. To improve the validity of test scores, greater weight should be given to fluency and lexical diversity in the listen-to-speak task, whereas lexical sophistication should be emphasized in the read-to-speak task. In instruction, teachers should help learners develop lexical fluency, sophistication, and diversity, but with a different focus at each proficiency level. For intermediate learners, the emphasis should shift toward lexical sophistication, encouraging them to use more difficult words when speaking.

The current study provides meaningful steps toward understanding the intricate role of vocabulary in integrated speaking assessments. To expand the generalizability of these findings, future research should incorporate larger, more balanced spoken corpora representing diverse cohorts of EFL learners. Furthermore, to empirically validate the textual scaffolding effects discussed herein, future studies need to trace how source text prompts are recycled in learners' oral output. Despite the limitations, the nuanced developmental trajectories and task-dependent variations identified in this study offer critical, empirical benchmarks. Ultimately, these insights would hold substantial promise for refining pedagogical interventions, optimizing integrated test designs, and informing the ongoing development of automated speech evaluation systems in the field of applied linguistics.

Abbreviations

AoA: Age of Acquisition

CEFR: Common European Framework of Reference for Languages

COCA: Corpus of Contemporary American English

C/S: Clauses per Sentence

DP: Dispersion (or Deviation of Proportions)

HD-D: Hypergeometric Distribution (applied to Type-Token Ratio)

KELSC: Korean English Learners' Spoken Corpus

LASSO: Least Absolute Shrinkage and Selection Operator

LR: Likelihood Ratio

MI^2: Mutual Information 2 score (collocation association metric)

MLC: Mean Length of Clause

MLTU: Mean Length of T-unit

MRC: Medical Research Council (Psycholinguistic Database)

MSTTR: Mean Segmental Type-Token Ratio

MTLD: Measure of Textual Lexical Diversity

PPOM: Partial Proportional Odds Modeling / Model

USF: University of South Florida (Free Association Norms)

VIF: Variance Inflation Factor

Declarations

Clinical trial number

Not applicable

Ethics approval and consent to participate

Not applicable. This study utilized the Korean English Learners' Spoken Corpus (KELSC), which is a publicly available, fully anonymized spoken corpus. Because the data were obtained from a public repository and contained no personally identifiable information, institutional review board (IRB) approval and informed consent to participate were not required.

Consent for publication

Not applicable. The manuscript does not contain any individual person's identifying details, images, or videos. As the sole author, I have reviewed and approved the final manuscript for publication.

Funding

This work was supported by the Kwangwoon University Research Grant in 2026.

Author Contribution

As the sole author, H.L. was responsible for the data analysis, manuscript drafting, and subsequent revisions.

Acknowledgement

This work was supported by the Kwangwoon University Research Grant in 2026.

Data Availability

The Korean English Learners' Spoken Corpus (KELSC) analyzed during the current study is publicly available in the Institute for Corpus Research, Incheon National University repository at http://icr.or.kr/miscellaneous/. The R codes utilized for text preprocessing and statistical analyses are available from the author upon request.

References

1. Abrams, Z. I. (2019). The effects of integrated writing on linguistic complexity in L2 writing and task-complexity. *System*, *81*, 110-121.
2. Appel, R., Trofimovich, P., Saito, K., Isaacs, T., & Webb, S. (2019). Lexical aspects of comprehensibility and nativeness from the perspective of native-speaking English raters. *ITL-International Journal of Applied Linguistics*, *170*(1), 24-52.
3. Ariamanesh, A. A., Barati, H., & Youhanaee, M. (2022). TOEFL iBT Iranian Test-Takers' Oral Language Performance: A Comparison between Independent and Integrated Speaking Tasks. *International TESOL Journal, 17*(1), 25-43.
4. Bestgen, Y. (2024). Back to Basics in Measuring Lexical Diversity: Too Simple to Be True. *Applied Linguistics*, amae053. https://doi.org/10.1093/applin/amae053
5. Bottini, R. (2022). *Lexical complexity in L2 English speech: Evidence from the Trinity Lancaster Corpus* [Doctoral dissertation, Lancaster University]. Lancaster University. https://doi.org/10.17635/lancaster/thesis/1605
6. Bottini, R., & Le Foll, E. (2025). The more proficient the learners, the less sophisticated their L2 vocabulary? The curious effect of the reference corpus on mean-frequency measures of lexical sophistication. *International Journal of Learner Corpus Research*, *11*(1), 47-78.
7. Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics, 46*(4), 1171–1178. https://doi.org/10.2307/2532457
8. Cai, H. (2020). Relating lexical and syntactic knowledge to academic English listening: The importance of construct representation. *Frontiers in Psychology*, *11*, 494.
9. Christensen, R. H. B. (2019). *ordinal—Regression models for ordinal data* (R package version 2019.12-10) [Computer software]. https://CRAN.R-project.org/package=ordinal
10. Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35*(2), 115-135.

11. Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*(2), 171-192.
12. Crossley, S., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly, 11*(3), 250-270.
13. Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, *48*(4), 1227-1237.
14. Crosthwaite, P. R. (2011). The effect of collaboration on the cohesion and coherence of L2 narrative discourse between English NS and Korean L2 English users. *Asian EFL Journal, 13*, 135–166.
15. Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal, 99*(1), 80-95.
16. Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*(1), 5-43.
17. De Jong, N. H., & Mora, J. C. (2019). Does having good articulatory skills lead to more fluent speech in first and second languages? *Studies in Second Language Acquisition*, *41*(1), 227-239.
18. Dóczi, B., & Kormos, J. (2015). *Longitudinal developments in vocabulary knowledge and lexical organization*. Oxford University Press.
19. Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of second language writing*, *4*(2), 139-155.
20. Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal*, *104*(2), 381-400.
21. Green, A. (2013). *Exploring language assessment and testing: Language in action*. Routledge.
22. Gregori-Signes, C., & Clavel-Arroitia, B. (2015). Analysing lexical density and lexical diversity in university students' written discourse. *Procedia-Social and Behavioral Sciences, 198*, 546-556.
23. Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218-238.
24. Ha, H. S. (2019). Lexical Richness in EFL Undergraduate Students' Academic Writing. *English Teaching*, *74*(3), 3-28.
25. Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.
26. Heng, R., Pu, L., & Liu, X. (2023). The effects of genre on the lexical richness of argumentative and expository writing by Chinese EFL learners. *Frontiers in Psychology*, *13*, 1082228.
27. Jung, C. K. (2021). Designing and building the Korean English Learners' Spoken Corpus (KELSC). *Studies in Foreign Language Education, 35*(3), 209-223.

28. Kao, S. M., & Wang, W. C. (2014). Lexical and organizational features in novice and experienced ELF presentations. *Journal of English as a Lingua Franca*, *3*(1), 49-79.
29. Khalifa, H., & Weir, C. J. (2009). *Examining reading*. Cambridge.
30. Kim, J. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English Teaching, 69*(4), 27-51.
31. Kim, M. (2024). The Relationships between Lexical Sophistication and Second Language Writing and Speaking. *English Language & Literature Teaching*, *30*(4), 27-48.
32. Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical Sophistication as a Multidimensional Phenomenon: Relations to Second Language Lexical Proficiency, Development, and Writing Quality. *The Modern Language Journal, 102*(1), 120–141.
33. Kimppa, L., Shtyrov, Y., Hut, S. C., Hedlund, L., Leminen, M., & Leminen, A. (2019). Acquisition of L2 morphology by adult language learners. *Cortex*, 116, 74-90.
34. Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System, 40*(4), 554–564.
35. Kolahi Ahari,M. , Ghonsooly,B. , Ghapanchi,Z. and Soodmand Afshar,H. (2025). Examining the Role of Lexical Sophistication, Lexical Diversity, Syntactic Sophistication, Syntactic Complexity, and Cohesion in L2 Speaking Proficiency Assessment. *International Journal of Language Testing, 15*(2), 43-70. doi: 10.22034/ijlt.2025.492133.1395
36. Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly, 49*(4), 757–786. https://doi.org/10.1002/tesq.194
37. Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12-24.
38. Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing, 33*(3), 319-340.
39. Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. The *Modern Language Journal, 102*(2), 333-349.
40. Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods 50*(3), 1030-1046. https://doi.org/10.3758/s13428-017-0924-4
41. Kyle, K. (2020). The relationship between features of source text use and integrated writing quality. *Assessing Writing, 45*, 100467.
42. Kyle, K., Sung, H., Eguchi, M., & Zenker, F. (2024). Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. *Studies in Second Language Acquisition, 46*(1), 278–299.
43. Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language learning*, *61*(2), 647-672.

44. Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English* (Lund Studies in English 74). Lund University.
45. Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, *15*(4), 474-496.
46. Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190-208.
47. Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. *Insights into non-native vocabulary teaching and learning*, *52*, 83-98.
48. Miralpeix, I., & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, *56*(1), 1-24.
49. Monteiro, K. R., Crossley, S. A., & Kyle, K. (2020). In search of new benchmarks: Using L2 lexical frequency and contextual diversity indices to assess second language writing. *Applied Linguistics*, *41*(2), 280-300.
50. Nation, I. S. P. (2013). *Learning Vocabulary in Another Language* (2nd ed.). Cambridge: Cambridge University Press.
51. O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language testing*, *12*(2), 217-237.
52. Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*. Cambridge university press.
53. Sung, H., Wolf, M. K., Suhan, M., & Kyle, K. (2025). Lexical richness in young English learners' writing: A focus on opinion and listen-write task types. *Assessing Writing, 66*, 100975. https://doi.org/10.1016/j.asw.2025.100975
54. Peterson, B., & Harrell, F. E., Jr. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 39*(2), 205–217. https://doi.org/10.2307/2347760
55. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
56. Tsunemoto, A., & Trofimovich, P. (2024). Coherence and comprehensibility in second language speakers' academic speaking performance. *Studies in Second Language Acquisition, 46*(3), 795-817.
57. Uchihara, T., & Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The language learning Journal*, *47*(1), 64-75.
58. Uchihara, T., & Clenton, J. (2023). The role of spoken vocabulary knowledge in second language speaking proficiency. *The Language Learning Journal*, *51*(3), 376-393.
59. Ure, J. (1971). Lexical density and register differentiation. *Applications of linguistics*, *23*(7), 443-452.
60. Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition, 42*(2), 383–410. https://doi.org/10.1017/S0272263119000676

61. Vasylets, O., Gilabert, R., & Manchón, R. M. (2017). The effects of mode and task complexity on second language production. *Language Learning*, *67*(2), 394-430.
62. Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.
63. Wurm, M. J., Rathouz, P. J., & Hanlon, B. M. (2021). Regularized ordinal regression and the ordinalNet R package. *Journal of Statistical Software, 99*(6), 1–42. https://doi.org/10.18637/jss.v099.i06
64. Yang, W., Cushing, S. T., & Yu, G. (2025). Linguistic predictors of L2 writing performance: Variations across genres. *Assessing Writing*, *66*, 100985.
65. Zhang, X., & Lu, X. (2014). A longitudinal study of receptive vocabulary breadth knowledge growth and vocabulary fluency development. *Applied Linguistics*, *35*(3), 283-304.